UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Chair of Practical Computer Science IV:
Dependable Systems Engineering

Master's Thesis

Vulnerabilities in Privacy-Preserving Record Linkage: The Threat of Dataset Extension Attacks

as part of the degree program Master of Science Business Informatics submitted by

Marcel Mildenberger
Matriculation number 1979905

on February 6, 2025.

Supervisor: Prof. Dr. Frederik Armknecht
PhD Student Jochen Schäfer

Abstract

The abstract should serve as an independent piece of information on your Thesis conveying a concise description of the main aspects and most important results. It should not be excessively long.

Write the abstract.

Contents

List of Figures

List of Tables

List of Algorithms

List of Code Snippets

Acronyms

BF Bloom Filter

DEA Dataset Extension Attack

GMA Graph Matching Attack

PII personally identifiable information

PPRL Privacy-Preserving Record Linkage

TMH Tabulation MinHash Encoding

TSH Two-Step Hash Encoding

1. Introduction

Data and record linkage is an important aspect of research and software projects, enabling the integration of data from different sources about the same entity to gain additional insights. This is particularly important in sectors such as healthcare or social sciences. In the United States, for example, the fragmented healthcare and public health ecosystem has benefited greatly from effective data linkage. The COVID-19 pandemic highlighted the critical importance of timely, accurate and efficient data linkage, as the lack of it led to problems in integrating disease and vaccination data. In response, projects have been launched by organisations such as the Centers for Disease Control and Prevention and the Food and Drug Administration to address these challenges and further develop linkage techniques. [**vidanage2024**]

In scenarios such as the COVID-19 pandemic, the entities linked during data integration are often natural persons. As a result, linking data from different sources, such as healthcare providers, typically requires the use of personally identifiable information (PII) as an identifier. However, the use of PII raises significant privacy concerns, as individuals could be identified and data breaches could have serious consequences. To mitigate these risks, techniques are required to protect PII by encrypting data prior to data linkage. [**vidanage2024**]

In order to still be able to perform linkage while preserving privacy, similarity preserving encodings are applied to the identifiers. Without such similiraty preserving encoding, matches between encrypted entities in different databases would not be possible. Over time, three main privacy-preserving encoding schemes have emerged as enablers for Privacy-Preserving Record Linkage (PPRL) [**vidanage2024**; SAH24]:

Bloom Filter (BF) Encoding, based on Bloom filters, is the most widely used and is considered the reference standard. They were introduced due to their simplicity and efficiency in terms of storage and computation of private set similarities. Further research also exists to improve BF by adding a diffusion layer to it. Tabulation MinHash Encoding (TMH), a more recent method with distinct advantages in certain use cases, is a variation of the MinHash algorithm. Even though it is less used in PPRL, due to its simplicity and efficiency in calculating set similarities its an viable option. It is generally more secure than BF encoding but requires more computation and memory. Two-Step Hash Encoding (TSH) provides a novel approach for encoding to solve the problems of BF and TMH. It uses a two-step hashing process to first encode the data in multiple BF and then apply a second hashing step to produce a set of integers used for similarity comparisons. [**vidanage2024**; SAH24]

In practice, BF based PPRL has become the dominant standard and is widely used in areas such as crime detection, fraud prevention and national security. However, PPRL has its limitations and vulnerabilities. As seen in previous research, several attacks on PPRL systems exists. These are attacks are soley focusing on BF encodings and therefore trying to exploit the encoding scheme itself. No Particular attack has been developed for TMH or TSH encodings. These attacks on BF are focusing on the weakness of the double hasing method, weaknesses of the BF construction orunciple, leveraging maximal frequent pattern-mining and language model or graph based dictionary attacks. [**vidanage2024**] Nevertheless, a more recent attack has emerged which leverage the vulnerabilities of PPRL systems in general making it applicable to all encoding schemes. The so called Graph Matching Attack (GMA), which exploits publicly available data to re-identify encrypted individuals based on overlapping records in a plaintext database like a phone book and the encrypted records, is this novel attack. [**vidanage2024**; SAH24] While GMAs can re-identify individuals present in both the plaintext and encrypted database no matter which encoding scheme was choosen by solving a graph isomorphism problem, their scope is limited to the overlap of the two datasets [SAH24].

This work aims to go beyond GMAs by re-identifying not only overlapping individuals, but as many individuals as possible from the encrypted PPRL data. To achieve this, the newly introduced Dataset Extension Attack (DEA) builds on GMAs. The DEA uses a neural network trained on previously decoded data to predict and re-identify the remaining encrypted records, significantly extending the scope and effectiveness of the attack.

1.1. Motivation

The increasing use of PPRL in highly sensitive areas requires further research to validate existing techniques and ensure robust data protection. While data privacy has always been a critical concern, its importance continues to grow in an era dominated by artificial intelligence and machine learning. As models increasingly rely on large datasets and data brokerage becomes more frequent, privacy protection has become a pressing issue.

As highlighted in the introduction, researchers have already demonstrated that PPRL systems are vulnerable to GMAs. These attacks, which allow re-identification of encrypted individuals, directly undermine the primary objective of PPRL. Although current GMAs are limited to overlapping data between encrypted and plaintext databases, the potential implementation of a DEA could introduce even greater risks. Such an attack would allow complete re-identification of encrypted databases, effectively nullifying the privacy guarantees of PPRL and rendering it useless.

The motivation for this work is to proactively demonstrate the consequences of such an attack in order to prevent it from happening in real-world scenarios. This research will expose potential vulnerabilities in PPRL systems by showing how attackers could exploit encrypted data. A successful implementation will demonstrate that state-of-the-art methods, such as Bloom filter-based PPRL, are not secure or robust enough for continued use. By highlighting these threats, this thesis aims to provide a basis for further research into the development of more secure and robust PPRL techniques.

1.2. Related Work

The work of Vidanage et al. introduces a novel attack against PPRL, known as the GMA. This attack exploits the similarity-preserving properties of commonly used encoding schemes, such as Bloom filters, making it universally applicable across various PPRL methods. By utilizing a graph-based approach, the GMA aligns nodes in similarity graphs to successfully re-identify individuals [**vidanage2024**]. This work is critical to the present study, as the DEA builds upon the re-identified individuals produced by the GMA.

Another key contribution comes from Schäfer et al., who revisit and extend the work of Vidanage et al. They provide a thorough reproduction and replication of the proposed GMA, uncovering an undocumented preprocessing step in the original codebase that unintentionally impacted the attack's success rate. This step was intended to improve performance but instead introduced implementation errors. Schäfer et al. addressed this issue by correcting the preprocessing and enhancing the GMA to improve both robustness and efficiency. Their improved implementation achieved higher re-identification rates compared to the original approach by Vidanage et al. [SAH24]. The work of Schäfer et al. is particularly relevant to this thesis, as their enhanced GMA implementation serves as the foundation for the DEA. The DEA relies on the re-identification of individuals from the GMA to further extend its capabilities and achieve broader de-anonymization.

1.3. Contribution

The contribution of this thesis is divided into three main parts. First, a comprehensive analysis of PPRL is performed, with a particular focus on the three major encoding schemes: Bloom Filter Encoding, Two-Step Hash Encoding and Tabulation MinHash Encoding. Next, the current state of the art GMA is analysed and its limitations are discussed in detail. However, the main focus of this thesis is the implementation and evaluation of the DEA, with the aim of decoding more individuals than is possible with the GMA.

To achieve this, the thesis will detail the conceptual foundations, requirements and theoretical underpinnings of the DE attack. Building on the results of the GMA, the DE attack will be implemented and adapted to maximise its effectiveness. The novel DE attack will then be evaluated against the three major PPRL encryption schemes. While the encoding schemes are less critical for the GMA due to its reliance on solving a graph isomorphism problem, they play a crucial role in the DE attack. This is because the neural network used in the DE attack must be trained specifically for each encryption scheme. However, the DE attack is designed to be adaptable to different encoding schemes, ensuring flexibility and applicability.

The core contribution of this thesis is to address the following research questions:

- How effective are supervised machine learning-based DEAs in re-identifying the remaining entries of an GMA?
- How do different encoding schemes affect the performance of a DEA?

1.4. Organization of this Thesis

This thesis is divided into X main sections. First, an overview of PPRL systems will be provided, with a particular focus on a thorough analysis of the most commonly used encoding techniques. Following this, the existing GMA will be introduced and explained to establish the foundation for the study. Additionally, an overview of neural networks will be presented to provide necessary background knowledge.

Next, a detailed description of the attack model for the DEA will be outlined, including how neural networks are leveraged to enhance the attack. This is followed by an explanation of the actual implementation of the DEA, along with a discussion of the experiments conducted. The results of the DEA across different encoding schemes will then be analyzed and evaluated. Finally, the thesis will conclude with a summary of the key contributions, a discussion of the broader implications, and suggestions for future research.

2. Background

2.1. Overview of PPRL

2.2. Key encoding techniques

2.2.1. Bloom Filters

2.2.2. Tabulation MinHash

2.2.3. Two-Step Hashing

2.3. Graph Matching Attacks

3. Methodology

3.1. Conceptual framework for Dataset Extension

3.2. Implementation

4. Results

4.1. Analysis

4.2. Discussion

5. Conclusion

5.1. Summary

5.2. Future Work

Bibliography

[PSZ+24] Aditi Pathak, Laina Serrer, Daniela Zapata, Raymond King, Lisa B. Mirel, Thomas Sukalac, Arunkumar Srinivasan, Patrick Baier, Meera Bhalla, Corinne David-Ferdon, Steven Luxenberg, and Adi V. Gundlapalli. "Privacy Preserving Record Linkage for Public Health Action: Opportunities and Challenges." In: *Journal of the American Medical Informatics Association* 31.11 (2024). Advance access publication July 24, 2024, pp. 2605–2612. DOI: 10.1093/jamia/ocae196.

[SAH24] Jochen Schäfer, Frederik Armknecht, and Youzhe Heng. "R+R: Revisiting Graph Matching Attacks on Privacy-Preserving Record Linkage." In: *Proceedings of [Conference Name, if available]*. Available at: https://github.com/SchaeferJ/graphMatching. University of Mannheim. 2024.

A. Auxiliary Information

Eidesstattliche Erklärung

Hiermit versichere ich, dass diese Abschlussarbeit von mir persönlich verfasst ist und dass ich keinerlei fremde Hilfe in Anspruch genommen habe. Ebenso versichere ich, dass diese Arbeit oder Teile daraus weder von mir selbst noch von anderen als Leistungsnachweise andernorts eingereicht wurden. Wörtliche oder sinngemäße Übernahmen aus anderen Schriften und Veröffentlichungen in gedruckter oder elektronischer Form sind gekennzeichnet. Sämtliche Sekundärliteratur und sonstige Quellen sind nachgewiesen und in der Bibliographie aufgeführt. Das Gleiche gilt für graphische Darstellungen und Bilder sowie für alle Internet-Quellen.
Ich bin ferner damit einverstanden, dass meine Arbeit zum Zwecke eines Plagiatsabgleichs in elektronischer Form anonymisiert versendet und gespeichert werden kann.

Datum

Marcel Mildenberger